

Introduction to Bayesian inference for astronomy

Tom Loredo

Cornell Center for Astrophysics and Planetary Science,
Carl Sagan Institute,
& Dept. of Statistics and Data Science, Cornell U.
<http://hosting.astro.cornell.edu/~loredo/>

CASt Summer School — 5–9 June 2023

For selected texts, tutorials, software on Bayesian data analysis for physicists, astronomers, and other scientists, see the *Resources* document in the [GitHub repo for this session](https://github.com/tloredo/SummerSchool2023-IntroBayes/):

<https://github.com/tloredo/SummerSchool2023-IntroBayes/>

Entry points for literature comparing Bayesian and frequentist approaches:

- Jaynes (1976): Confidence Intervals vs Bayesian Intervals (article # 32)
- Loredó (1992): The promise of Bayesian inference for astrophysics; unabridged version at BIPS
- Loredó (2013): Bayesian astrostatistics: A backward look to the future
- Diaconis & Skyrms (2017): *Ten Great Ideas about Chance* (See *Resources*)

Scientific method

*Science is more than a body of knowledge; it is a way of thinking.
The method of science, as stodgy and grumpy as it may seem,
is far more important than the findings of science.*
—Carl Sagan

Scientists *argue!*

Argument \equiv Collection of statements comprising an act of reasoning from *premises* to a *conclusion*

A key goal of science: Explain or predict *quantitative measurements* (data!)

Data analysis: Constructing and appraising arguments that reason from data to interesting scientific conclusions (explanations, predictions)

The role of data

Data do not speak for themselves!

*“No body of data tells us all we need to know
about its own analysis.”*

— John Tukey, *EDA*

We don't just *tabulate* data, we *analyze* data

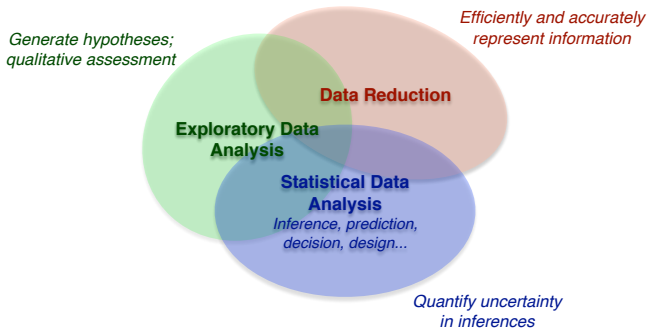
We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses

A key role of data in science is to be among the premises in scientific arguments—the evidence for evidence-based reasoning

Data analysis

Building & Appraising Arguments Using Data

Modes of Data Analysis



Inference: Learning about the *data generating process* (population, signals. . .) from observed data—just one of several interacting modes of analyzing data

Statistical inference as a style of reasoning

For better or worse, statistical inference has provided an entirely new *style of reasoning*. The quiet statisticians have changed our world—not by discovering new facts or technical developments but by changing the ways we reason, experiment, and form our opinions about it.

—*Ian Hacking, philosopher of science (in Science '84)*

Two main styles of statistical reasoning

Statistics has not yet aged into a stable discipline with complete agreement on foundations. All the statisticians mentioned here [Pearson, Galton, Fisher] assumed that *the key to probability lay in the relative frequency with which different kinds of events occur*. . . . But what does that mean? Some say nothing, for probability is concerned with subjective degrees of belief, and that subjective approach only gives a *reasonable degree of certainty*. Work emanating from F. P. Ramsey in England, Bruno de Finetti in Italy, and L. J. Savage in the United States has turned such subjectivity into a serious scientific approach. Today we have vigorous, sometimes violent, disagreement on these matters, but perhaps battles about first principles are less important than the large-scale application of many competing methods.

—*Ian Hacking (1984)*

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Fundamental principle

*“The most fundamental principle of the statistical paradigm,
its starting point,
is that variation may be described by probability.”*

Fundamental principle

~~*“The most fundamental principle of the statistical paradigm,
its starting point,
is that variation may be described by probability.”*~~

Fundamental principle

*The most fundamental principle of the statistical paradigm,
its starting point,
is that **uncertainty** may be described by probability.*

Fundamental principle

*The most fundamental principle of the statistical paradigm,
its starting point,
is that **uncertainty** may be described by probability.*

*An important corollary is that, in some settings
—most notably, for **IID replications**,
and for **exchangeable sequences**—
expected variation and individual-case probability
are intimately linked.*

Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation*.

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful*, none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities*, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Harold Jeffreys (1931)

If we like there is no harm in saying that a *probability expresses a degree of reasonable belief*. . . . 'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. *Essentially the notion can only be described by reference to instances where it is used*. It is intended to express *a kind of relation between data and consequence* that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

Probability theory as (generalized) logic

“Logic can be defined as *the analysis and appraisal of arguments*”
—Gensler, *Intro to Logic*

Build arguments with *propositions* and logical
operators/connectives

- *Propositions*: Statements that may be true or false

\mathcal{P} : Universe can be modeled with Λ CDM

A : $\Omega_{\text{tot}} \in [0.9, 1.1]$

B : Ω_{Λ} is not 0

\overline{B} : “not B ,” i.e., $\Omega_{\Lambda} = 0$

Events in freq. PT are propositions about outcomes in repeated trials

- *Connectives*:

$A \wedge B$ or A, B : A and B are both true

$A \vee B$: A or B is true, or both are

Arguments

Argument: Assertion that an *hypothesized conclusion*, H , follows from *premises*, $\mathcal{P} = \{A, B, C, \dots\}$ (take “,” = “and”)

Notation:

$H|\mathcal{P}$: Premises \mathcal{P} imply H
 H may be deduced from \mathcal{P}
 H follows from \mathcal{P}
 H is true given that \mathcal{P} is true

Deductive logic applies when we can *reason with certainty*; can model this with *Boolean algebra* over $\{0, 1\}$ (False, True)

Classical/Bayesian PT applies when we must *reason amidst uncertainty*; it quantifies degree of certainty on a $[0, 1]$ scale, providing a mathematical model for *inductive reasoning*

Probability as argument strength

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

$P = 1 \rightarrow$ Argument is *deductively valid*

$= 0 \rightarrow$ Premises imply \bar{H}

$\in (0, 1) \rightarrow$ Degree of deducibility

Mathematical model for induction

$$\begin{aligned}\text{'AND' (product rule): } P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P}) P(B|A \wedge \mathcal{P}) \\ &= P(B|\mathcal{P}) P(A|B \wedge \mathcal{P})\end{aligned}$$

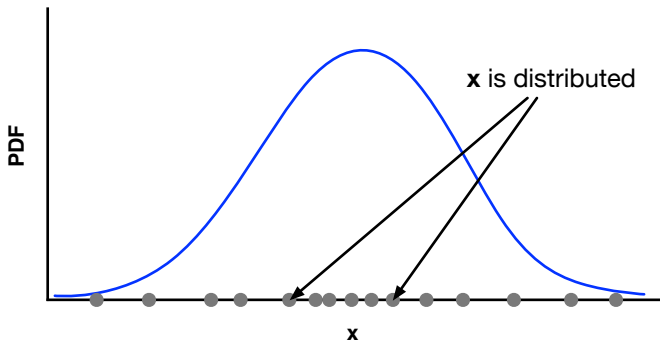
$$\begin{aligned}\text{'OR' (sum rule): } P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\ &\quad - P(A \wedge B|\mathcal{P})\end{aligned}$$

$$\text{'NOT': } P(\bar{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

Interpreting PDFs

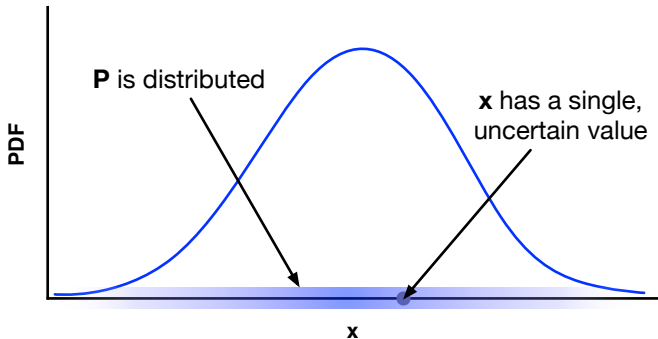
Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. $p(x)$ describes how the *values of x* would be distributed among *infinitely many trials*:



Bayesian

Probability *quantifies uncertainty* in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in *the single case before us*:



Probability & frequency in IID settings

Consider a setting where we assign the same probability to many independent outcomes (flips of a coin, rolls of a die, searches for an Earth around a G dwarf. . .):

- If the probability is high, we expect the outcomes to occur frequently
- If the probability is low, we expect the outcomes to occur rarely

In IID repeated trial settings, it seems there should be a relationship between single-trial probability and multiple-trial (relative) frequency

Frequency from probability

Bernoulli's (weak) law of large numbers: In repeated IID trials, given $P(\text{success}|\dots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \rightarrow \alpha \quad \text{as} \quad N_{\text{total}} \rightarrow \infty$$

If $P(\text{success}|\dots)$ does not change from sample to sample, it may be interpreted as the expected relative frequency

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" \rightarrow First use of Bayes's theorem:

Probability for success in next trial of IID sequence:

$$E(\alpha) \rightarrow \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \rightarrow \infty$$

If $P(\text{success}|\dots)$ does not change from sample to sample, it may be estimated using relative frequency data

Twiddle notation for the normal distribution

$$\text{Norm}(x; \mu, \sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{\sigma^2} \right]$$

Frequentist

random \searrow \swarrow fixed but unknown

$$p(x; \mu, \sigma) = \text{Norm}(x; \mu, \sigma)$$
$$x \sim \mathcal{N}(\mu, \sigma^2)$$

“ x is distributed as normal with mean...”

“random” = “varies unpredictably in repeated trials”

Bayesian

random \searrow \swarrow random or known

$$p(x | \mu, \sigma) = \text{Norm}(x | \mu, \sigma)$$
$$x \sim \mathcal{N}(\mu, \sigma^2)$$

“The probability for x is distributed as normal with mean...”

“random” = “uncertain in the case at hand”

Bayesian statistical inference

- Bayesian inference uses probability theory to *quantify the strength of data-based arguments* (i.e., a more abstract view than restricting PT to describe variability in repeated “random” experiments)
- A different approach to *all* statistical inference problems (i.e., not just another method in the list: BLUE, linear regression, least squares/ χ^2 minimization, maximum likelihood, ANOVA, survival analysis, LDA classification . . .)
- Focuses on *deriving consequences of modeling assumptions* rather than *devising and calibrating procedures*

Bayesian data analysis (BDA): Using Bayesian ideas across various data analysis tasks—not just inference, but also prediction, decision, design, EDA, data reduction. . .

Frequentist vs. Bayesian statements

“The data D_{obs} support hypothesis H . . . ”

Frequentist assessment

“ H was selected with a procedure that’s right 95% of the time over a set $\{D_{\text{hyp}}\}$ that includes D_{obs} .”

Probabilities are properties of *procedures*, not of particular results. Guaranteed long-run performance is the *sine qua non*.

Bayesian assessment

“The strength of the chain of reasoning from the model and D_{obs} to H is 0.95, on a scale where 1= certainty.”

Probabilities are associated with arguments based on *specific, observed data*.

Long-run performance must be separately evaluated (and is typically good by frequentist criteria).

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

A Simple (?) confidence region

Problem

Estimate the location (mean, μ) of a Gaussian distribution from a set of N IID samples $D = \{x_i\}$. Report a region summarizing the uncertainty.

Here assume std dev'n σ is *known*; we are uncertain only about μ

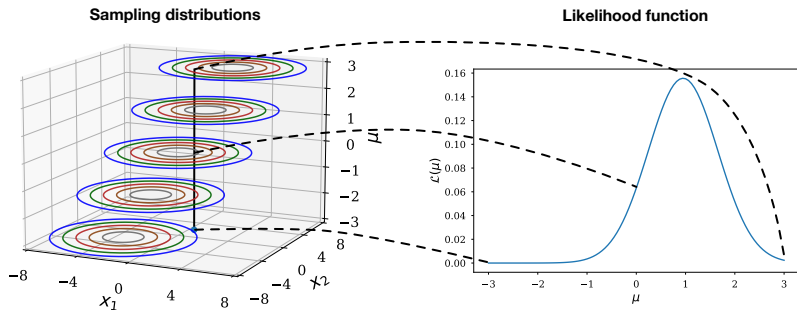
Model

The *sampling distribution* for *any* set $\{x_i\}$ is

$$\begin{aligned} p(\{x_i\}|\mu) &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; & \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

This gives the *likelihood function*, $\mathcal{L}(\mu)$ if we set $\{x_i\}$ to the *observed values*

Sampling distributions and likelihood function



The likelihood function shows how well each of the candidate sampling distributions—labeled by the parameter, μ —predicts the observed data (x_1, x_2) ; “predictive” or “prognostic” might have been better

The likelihood for the parameter is the (sampling) probability for the observed data; “likelihood for the data” is incorrect usage—it entirely misses the point of likelihood!

Fisher on likelihood

“If we need a word to characterise this relative property of different values of p , I suggest that we may speak without confusion of the likelihood of one value of p being thrice the likelihood of another, bearing always in mind that *likelihood is not here used loosely as a synonym of probability*, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample.” (Fisher 1922)

“Likelihood also *differs from probability* in that it is a differential element, and is *incapable of being integrated*: it is assigned to a particular point of the range of variation, not to a particular element [interval].” (Fisher 1922)

“... the integration with respect to m is illegitimate and has no definite meaning...” (Fisher 1912)

Classes of variables—the two spaces

- μ is the unknown we seek to estimate—the *parameter*. The *parameter space* is the space of possible values of μ —here the real line (perhaps bounded). *Hypothesis space* is a more general term.
- A particular set of N data values $D = \vec{x} = (x_1, \dots, x_N)$ is a *sample*. The *sample space* is the N -dimensional space of possible samples. The *observed* data correspond to a single point in this space.

Roles for probability

- Both frequentist and Bayesian approaches require probability distributions on the sample space—*sampling distributions*
- The frequentist perspective denies the legitimacy of probability distributions on the parameter space (the parameter value isn't random in the sense of stochastically varying). Inference is handled by describing the long-run performance of statistical procedures that produce data-based statements about parameters.
- The Bayesian approach allows probability distributions over parameter space, computing *posterior distributions* for inference.

Standard inferences for a normal mean

Inference takes us from the sample space to the parameter space.

Let $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$.

Point estimator: $\hat{\mu}(\vec{x}) = \bar{x}$.

Uncertainty quantification:

- “Standard error” (rms error) is σ/\sqrt{N}
- “1 σ ” interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level CL = 68.3%
- “2 σ ” interval: $\bar{x} \pm 2\sigma/\sqrt{N}$ with CL = 95.4%

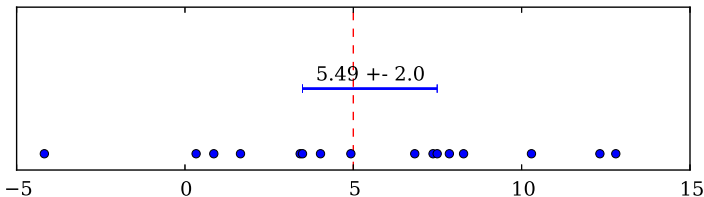
How are the uncertainty quantification results found?

Take a Monte Carlo perspective—for both frequentist & Bayesian approaches.

Some simulated data

Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

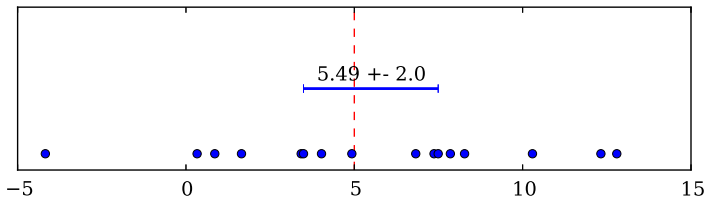
What is the CL associated with this interval?



Some simulated data

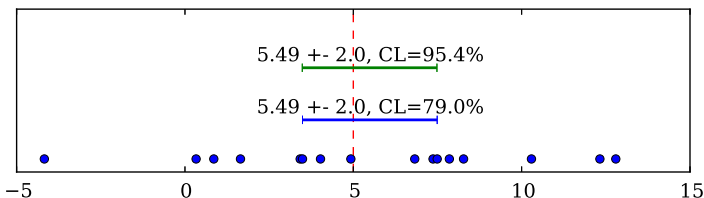
Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

What is the CL associated with this interval?



The confidence level for this interval is 79.0%.

Two intervals



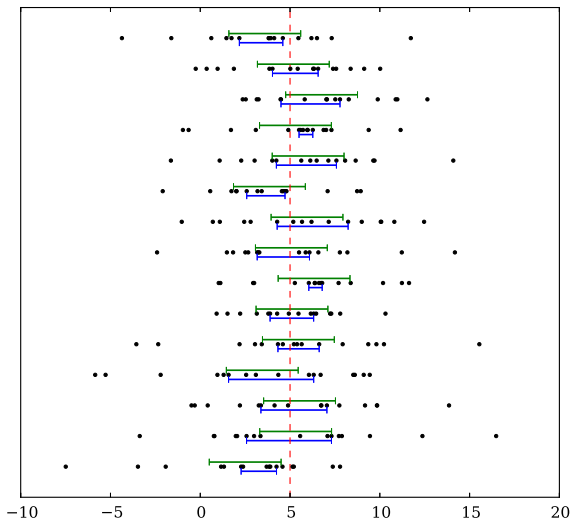
- Green interval: $\bar{x} \pm 2\sigma/\sqrt{N}$
- Blue interval: Let $x_{(k)} \equiv k$ 'th order statistic
Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

The point

*The (frequentist) confidence level is a **property of the procedure**, not of the particular interval reported for a given dataset*

Performance of intervals

Intervals for 15 datasets



Confidence interval for a normal mean

Suppose we have a sample of $N = 5$ values x_i ,

$$x_i \sim N(\mu, 1)$$

We want to estimate μ , including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/ χ^2 , maximum likelihood

Focus on likelihood (equivalent to χ^2 here); this is closest to Bayes:

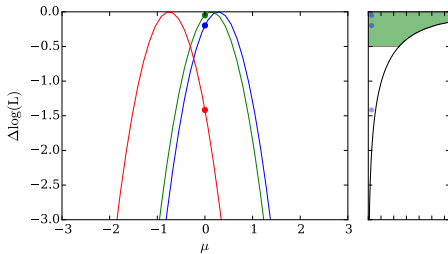
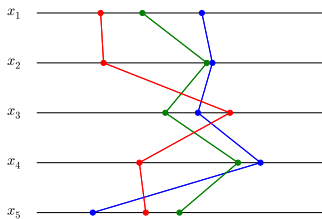
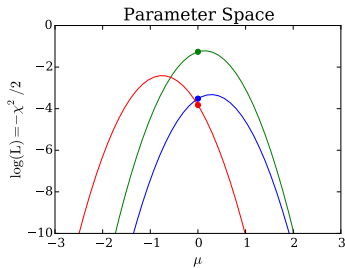
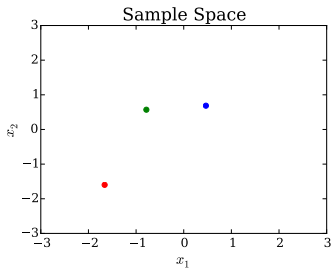
$$\begin{aligned}\mathcal{L}(\mu) &\equiv p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \quad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2}\end{aligned}$$

Estimate μ from maximum likelihood (minimum χ^2)

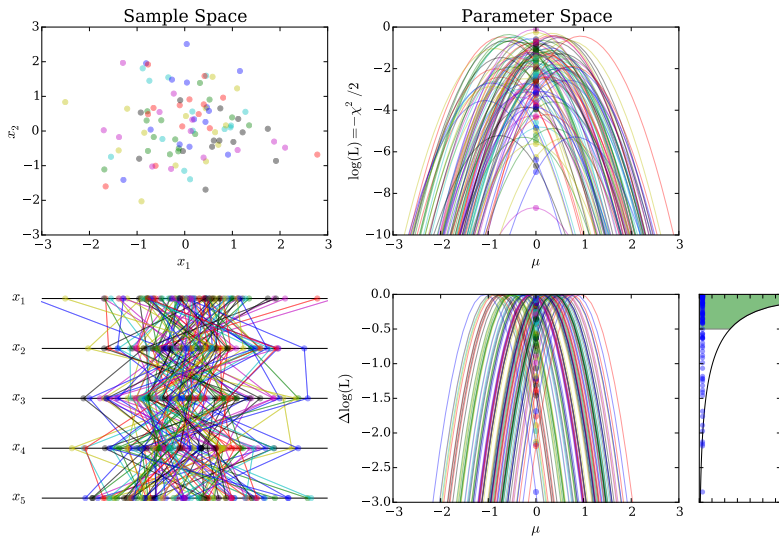
Define an interval and its coverage frequency from the $\mathcal{L}(\mu)$ curve

Construct an interval procedure for known μ

Likelihoods for 3 simulated data sets, $\mu = 0$

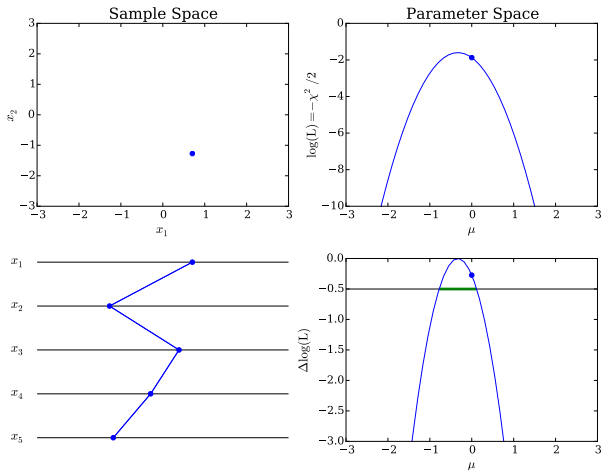


Likelihoods for 100 simulated data sets, $\mu = 0$



Careful! This is for $\mu = 0$, but μ will be unknown.
 Luckily, the $\Delta \log(\mathcal{L})$ dist'n is independent of μ .

Apply to observed sample



Report the green region, reporting CL as the coverage calculated for ensemble of hypothetical data (green region, previous slide)

Credible interval for a normal mean

Recall the likelihood, $\mathcal{L}(\mu) \equiv p(D_{\text{obs}}|\mu)$, is a probability for the observed data, but *not* for the parameter μ (wrong PDF units)

Convert likelihood to a probability distribution over μ via *Bayes's theorem* (changes units from D to μ):

$$\begin{aligned} p(\mu, D) &= p(\mu)p(D|\mu) \\ &= p(D)p(\mu|D) \\ \rightarrow p(\mu|D) &= p(\mu)\frac{p(D|\mu)}{p(D)}, \quad \text{Bayes's th.} \end{aligned}$$

$$\Rightarrow p(\mu|D_{\text{obs}}) \propto \pi(\mu)\mathcal{L}(\mu) \quad (\text{prior} \times \text{like.})$$

$p(\mu|D_{\text{obs}})$ is called the *posterior probability distribution for μ*

This requires a prior probability density, $\pi(\mu)$, often taken to be constant over the allowed region if there is no significant information available (or sometimes constant w.r.t. some reparameterization motivated by a symmetry in the problem)

Gaussian problem posterior distribution

For the Gaussian example, a bit of algebra (“complete the square”) gives:

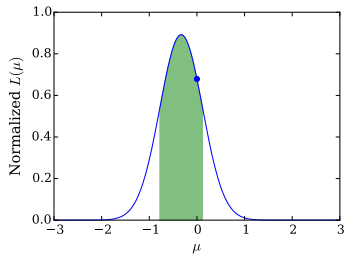
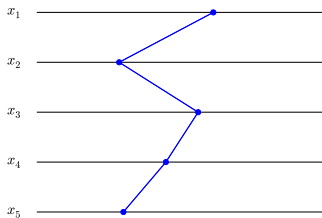
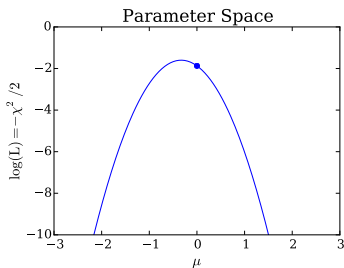
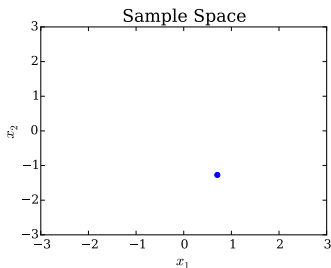
$$\begin{aligned}\mathcal{L}(\mu) &\propto \prod_i \exp \left[-\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &\propto \exp \left[-\frac{1}{2} \sum_i \frac{(x_i - \mu)^2}{\sigma^2} \right] \\ &\propto \exp \left[-\frac{(\mu - \bar{x})^2}{2(\sigma/\sqrt{N})^2} \right]\end{aligned}$$

The likelihood is Gaussian in μ .

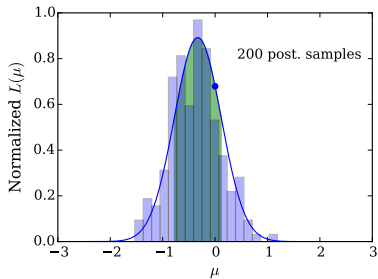
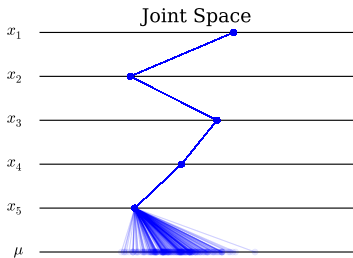
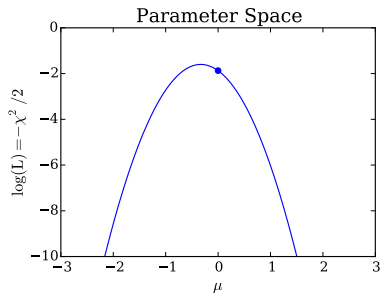
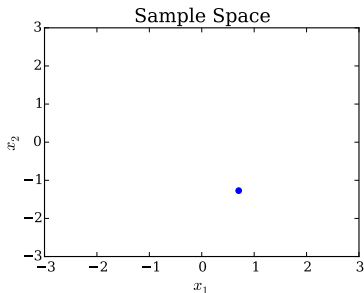
Flat prior \rightarrow posterior density for μ is $\mathcal{N}(\bar{x}, \sigma^2/N)$.

Bayesian credible region

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood; green shows the *highest posterior density (HPD) region*:



Posterior sampling: Credible region via Monte Carlo (MCMC, ABC)



Posterior summaries

- Posterior mean is $\langle \mu \rangle \equiv \int d\mu \mu p(\mu|D_{\text{obs}}) = \bar{x}$
- Posterior mode is $\hat{\mu} = \bar{x}$
- Posterior std dev'n is σ/\sqrt{N}
- $\bar{x} \pm \sigma/\sqrt{N}$ is a 68.3% *credible region*:

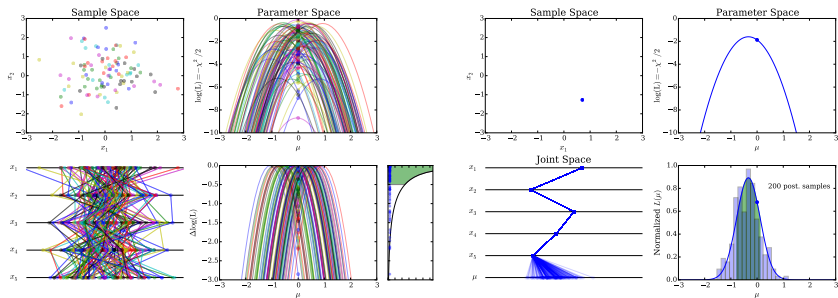
$$\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\mu p(\mu|D_{\text{obs}}) \approx 0.683$$

- $\bar{x} \pm 2\sigma/\sqrt{N}$ is a 95.4% credible region

The credible regions above are *highest posterior density* credible regions (HPD regions). These are the smallest regions with a specified probability content.

These reproduce familiar frequentist results, but this is a *coincidence* due to special properties of Gaussians.

Confidence vs. credible regions



Find lower/upper functions of the data that give desired values to:

$$\text{Cover}(\mu) = \int d^N \vec{x} \, p(\vec{x}|\mu) \, \mathbb{I} [l(\vec{x}) < \mu < u(\vec{x})]$$

$$\text{CredLev}(\vec{x}_{\text{obs}}) = \int d\mu \, p(\mu|\vec{x}_{\text{obs}}) \, \mathbb{I} [L(\vec{x}_{\text{obs}}) < \mu < U(\vec{x}_{\text{obs}})]$$

When the approaches differ

Both approaches report $\mu \in [\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}]$, and assign 68.3% to this interval (*with very different meanings!*)

This matching is a *coincidence*!

When might results differ? (\mathcal{F} = frequentist, \mathcal{B} = Bayes)

- If \mathcal{F} procedure doesn't use likelihood directly
- If \mathcal{F} procedure properties depend on params (e.g., nonlinear models; need to find pivotal quantities)
- If likelihood shape varies strongly between datasets (conditional inference, ancillary statistics, recognizable subsets)
- If there are extra uninteresting parameters (nuisance parameters; adjusted profile likelihood, conditional inference)
- If \mathcal{B} uses important prior information

Also, for a different task—comparison of parametric models—the approaches are *qualitatively* different (significance tests & info criteria vs. Bayes factors)

Bayesian and Frequentist inference

Brad Efron, ASA President (2005)

The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having *important practical consequences*... The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there's only one way of doing physics but there seems to be at least two ways to do statistics, and *they don't always give the same answers*...

Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist. What's going to happen in the 21st Century?... I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning...

Roderick Little, ASA President's Address (2005)

Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician's toolkit. . . . I am discomforted by this "inferential schizophrenia." Since the Bayesian (B) and frequentist (F) philosophies *can differ even on simple problems*, at some point decisions seem needed as to which is right. I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject. . . .

An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that *calibrated Bayes*. . . captures the strengths of both approaches and provides a roadmap for future advances.

[Calibrated Bayes = Bayesian inference within a specified space of models + frequentist-based model checking; Andrew Gelman et al. use *Bayesian data analysis* similarly]

(see TL's arXiv:1208.3035 for discussion/references)

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

The Bayesian recipe

Assess hypotheses by calculating their probabilities $p(H_i | \dots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises

$$\text{'OR' (sum rule): } P(H_1 \vee H_2 | \mathcal{C}) = P(H_1 | \mathcal{C}) + P(H_2 | \mathcal{C}) - P(H_1, H_2 | \mathcal{C})$$

$$\begin{aligned} \text{'AND' (product rule): } P(H_i, D_{\text{obs}} | \mathcal{C}) &= P(H_i | \mathcal{C}) P(D_{\text{obs}} | H_i, \mathcal{C}) \\ &= P(D_{\text{obs}} | \mathcal{C}) P(H_i | D_{\text{obs}}, \mathcal{C}) \end{aligned}$$

$$\text{'NOT': } P(\overline{H_i} | \mathcal{C}) = 1 - P(H_i | \mathcal{C})$$

Two Important Theorems

Bayes's Theorem (BT)

Consider the *joint probability* for a hypothesis and the observed data, $P(H_i, D_{\text{obs}}|\mathcal{C})$, using the product rule:

$$\begin{aligned}P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C}) P(D_{\text{obs}}|H_i, \mathcal{C}) \\&= P(D_{\text{obs}}|\mathcal{C}) P(H_i|D_{\text{obs}}, \mathcal{C})\end{aligned}$$

Solve for the *posterior probability* for H_i (adds a premise!):

$$P(H_i|D_{\text{obs}}, \mathcal{C}) = \frac{P(H_i, D_{\text{obs}}|\mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})} = P(H_i|\mathcal{C}) \frac{P(D_{\text{obs}}|H_i, \mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

$$\text{posterior} \propto \text{prior} \times \text{likelihood} \\ (\text{all "for } H_i\text{"})$$

$$\text{norm. const. } P(D_{\text{obs}}|\mathcal{C}) = \text{prior predictive for } D_{\text{obs}}$$

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ (“suite;” \mathcal{C} asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i | \mathcal{C}) &= \sum_i P(B_i | A, \mathcal{C}) P(A | \mathcal{C}) = P(A | \mathcal{C}) \\ &= \sum_i P(B_i | \mathcal{C}) P(A | B_i, \mathcal{C})\end{aligned}$$

If we do not see how to get $P(A | \mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a “basis”—*extend the conversation*:

$$P(A | \mathcal{C}) = \sum_i P(B_i | \mathcal{C}) P(A | B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A | \mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A | \mathcal{C}) = \sum_i P(A, B_i | \mathcal{C})$$

Joseph Blitzstein (Harvard statistician) on LTP (paraphrased):

In most areas of math, when you're stuck, saying, "I wish I knew this or that" doesn't help you. In probability theory, saying "I wish I knew this" suggests what to condition on; then you condition on it, compute *as if* you knew it, and then average over those possibilities.

*I didn't name the law of total probability, but if I had, I would have just called it **wishful thinking**.*

— YouTube lecture on conditional probability (15:48)

LTP example 1: Take \mathcal{C} to specify fair roll of a die, $A =$ “An even number comes up,” $B_i =$ “face i comes up” ($i = 1$ to 6)

$$\begin{aligned}P(A|\mathcal{C}) &= \sum_{i=1}^6 P(A, B_i|\mathcal{C}) \\&= \sum_{i=1}^6 P(B_i|\mathcal{C})P(A|B_i, \mathcal{C}) \\&= \frac{1}{6} \times (0 + 1 + 0 + 1 + 0 + 1) = \frac{1}{2}\end{aligned}$$

LTP example 2: With context \mathcal{C} , take $A = D_{\text{obs}}$, $B_i = H_i$; then

$$\begin{aligned}P(D_{\text{obs}}|\mathcal{C}) &= \sum_i P(D_{\text{obs}}, H_i|\mathcal{C}) \\&= \sum_i P(H_i|\mathcal{C})P(D_{\text{obs}}|H_i, \mathcal{C})\end{aligned}$$

prior predictive for $D_{\text{obs}} =$ Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Tabular/diagrammatic Bayesian inference

Simplest case: *Binary classification*

- 2 hypotheses: $\{C, \overline{C}\}$
- 2 possible data values: $\{-, +\}$

Concrete example: You test positive (+) for a medical condition. Do you have the condition (C) or not (\overline{C})?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
 - Test is 80% accurate if you have the condition:
 $P(+|C, \mathcal{C}) = 0.8$ (“sensitivity”)
 - Test is 95% accurate if you are healthy:
 $P(-|\overline{C}, \mathcal{C}) = 0.95$ (“specificity,” $1 - p(\text{false } +)$)

Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women

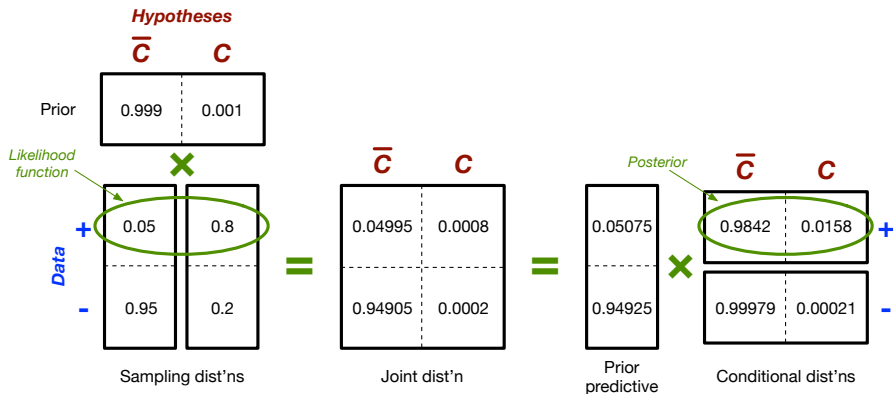
Tabular calculation

| Hypothesis H_i | Prior $\pi_i \equiv p(H_i)$ | Likelihood $\mathcal{L}_i \equiv p(+ H_i)$ | Joint $\pi_i \times \mathcal{L}_i$ | Posterior $p(H_i +)$ |
|---------------------|--------------------------------|---|---------------------------------------|-------------------------|
| \overline{C} | 0.999 | 0.05 | 0.04995 | 0.9842 |
| C | 0.001 | 0.8 | 0.0008 | 0.0158 |
| Sums: | 1.0 | NA | 0.05075 $= p(+)$ | 1.0 |

Inference as manipulation of the joint distribution

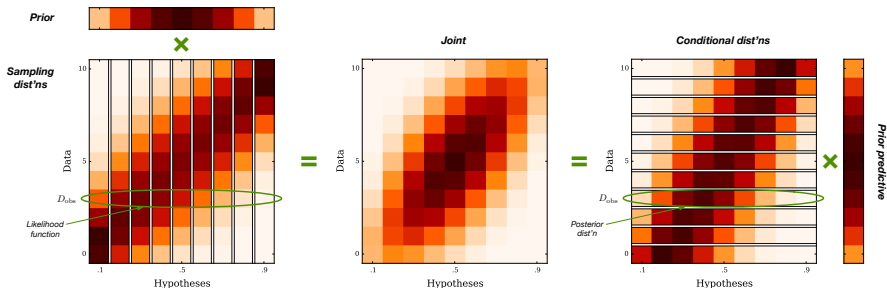
Bayes's theorem in terms of the *joint distribution*:

$$P(H_i|\mathcal{C}) \times P(D_{\text{obs}}|H_i, \mathcal{C}) = P(H_i, D_{\text{obs}}|\mathcal{C}) = P(H_i|D_{\text{obs}}, \mathcal{C}) \times P(D_{\text{obs}}|\mathcal{C})$$

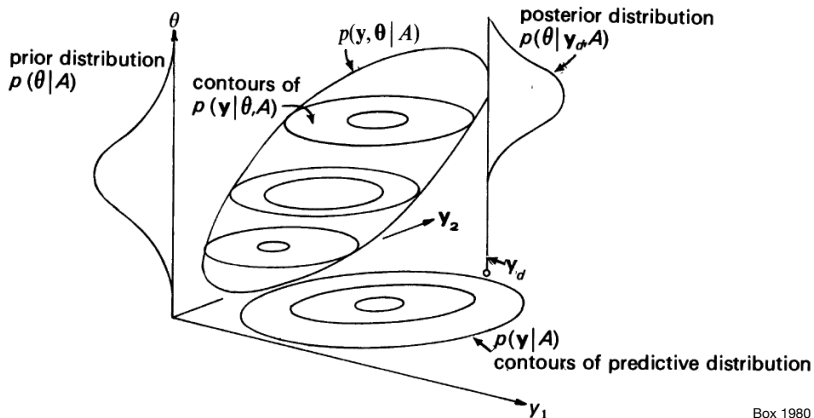


Larger discrete case: Flip a coin 10 times (Binomial inference)

- 9 hypotheses: Prob. for heads is $\alpha = 0.1, 0.2, \dots, 0.9$
- 11 possible data values: Number of heads, $n = 0, 1, \dots, 10$
- Adopt a prior concentrated around $\alpha = 0.5$, with some spread



Continuous data, parameter spaces



Box 1980

Components of Bayes's theorem for a problem with a 1-D parameter space (θ) and a 2-D sample space (\mathbf{y}), with observed data \mathbf{y}_d , and modeling assumptions A

Recap of key ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis \propto ability of hypothesis to *predict* the observed data

Law of total probability

$$p(\text{Hypotheseses} \mid \text{Data}) = \sum p(\text{Hypothesisis} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Inference with parametric models

Models M_i ($i = 1$ to N), each with parameters θ_i , each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty).

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{\text{obs}}$.

Classes of problems

Single-model inference

Premise = choice of single model (specific i)

Parameter estimation: What can we say about θ_i or $f(\theta_i)$?

Prediction: What can we say about future data D' ?

Multi-model inference

Premise = $\{M_i\}$

Model comparison/choice: What can we say about i ?

Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D' , accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ “all” alternatives

Is M_1 adequate? (predictive tests, calibration, robustness)

Parameter estimation

Problem statement

\mathcal{C} = Model M with parameters θ (+ any add'l info)

H_i = statements about θ ; e.g. " $\theta \in [2.5, 3.5]$," or " $\theta > 0$ "

Probability for any such statement can be found using a *probability density function* (PDF) for θ :

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \dots) &= f(\theta) d\theta \\ &= p(\theta | \dots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M) \mathcal{L}(\theta)}{\int d\theta p(\theta | M) \mathcal{L}(\theta)}$$

Summaries of posterior

- “Best fit” values:
 - ▶ Mode, $\hat{\theta}$, maximizes $p(\theta|D, M)$
 - ▶ Posterior mean, $\langle \theta \rangle = \int d\theta \theta p(\theta|D, M)$
- Uncertainties:
 - ▶ Credible region Δ of probability C :
 $C = P(\theta \in \Delta|D, M) = \int_{\Delta} d\theta p(\theta|D, M)$
Highest Posterior Density (HPD) region has $p(\theta|D, M)$ higher inside than outside
 - ▶ Posterior standard deviation, variance, covariances
- Marginal distributions
 - ▶ Interesting parameters ϕ , nuisance parameters η
 - ▶ Marginal dist'n for ϕ : $p(\phi|D, M) = \int d\eta p(\phi, \eta|D, M)$

Estimating a normal mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is known $\rightarrow I = (\sigma, M)$.

Parameter space: μ ; seek $p(\mu|D, \sigma, M)$

Likelihood

$$\begin{aligned}\mathcal{L}(\mu) &\equiv p(D|\mu, \sigma, M) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(d_i - \mu)^2 / 2\sigma^2}; \quad \sigma = 1 \\ &\propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right)\end{aligned}$$

Likelihood function is a Gaussian function at \bar{d} , width $w = \sigma/\sqrt{N}$

Informative conjugate prior

Use a normal prior, $\mu \sim \mathcal{N}(\mu_0, w_0^2)$

Conjugate because the posterior turns out also to be normal

Posterior

Normal $\mathcal{N}(\tilde{\mu}, \tilde{w}^2)$, but mean shifts towards prior, std. deviation decreases (reflecting add'l info from the prior)

Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large; then

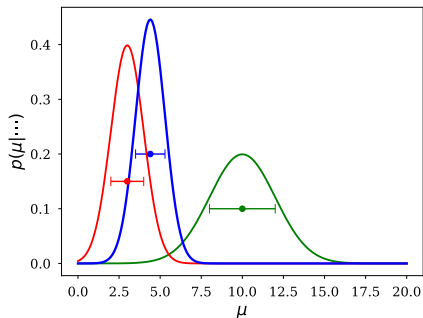
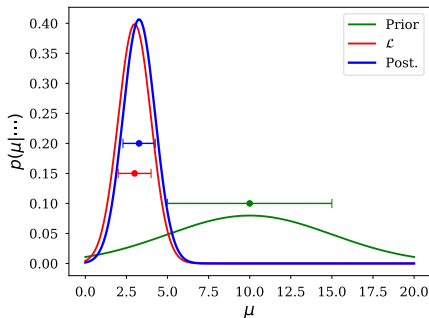
$$\begin{aligned}\tilde{\mu} &= \bar{d} + B \cdot (\mu_0 - \bar{d}) \\ \tilde{w} &= w \cdot \sqrt{1 - B}\end{aligned}$$

Principle of stable estimation/precise measurement — “If observations are precise. . . relative to the prior, then the form and properties of the prior distribution have negligible influence on the posterior distribution.”

Edwards, Lindman, and Savage (1963), ‘Bayesian Statistical Inference for Psychological Research,’ reprinted in *Breakthroughs in Statistics*

Conjugate normal examples:

- Data have $\bar{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$



Supplement:

- Binomial example
 - ▶ Bernoulli trials: Bernoulli process & binomial sampling dist'ns
 - ▶ Beta-binomial conjugate model
- Normal example, cont'd
 - ▶ Analytical details for normal example
 - ▶ Sufficiency; sample mean and variance as sufficient statistics
 - ▶ Handling σ uncertainty by marginalizing over σ ; Student's t distribution

Nuisance parameters and marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b .

We have additional data just about b .

What do the data tell us about s ?

Marginal posterior distribution

To summarize implications for s , accounting for b uncertainty, *marginalize*:

$$\begin{aligned} p(s|D, M) &= \int db \, p(s, b|D, M) \\ &\propto p(s|M) \int db \, p(b|s, M) \mathcal{L}(s, b) \\ &= p(s|M) \mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function* for s :

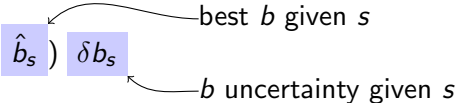
$$\mathcal{L}_m(s) \equiv \int db \, p(b|s) \mathcal{L}(s, b)$$

Maximum likelihood suggests instead computing the *profile likelihood*:

$$\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s), \quad \hat{b}_s = \text{best } b \text{ given } s$$

Marginalization vs. profiling

For insight: Suppose the prior is broad compared to the likelihood
→ for a fixed s , we can accurately estimate b with max likelihood \hat{b}_s , with small uncertainty δb_s .

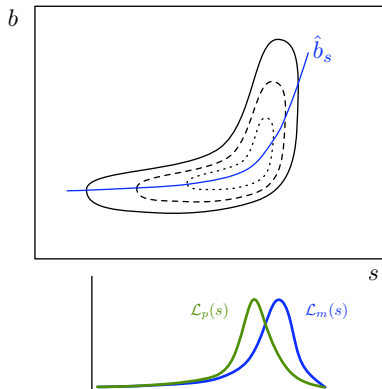
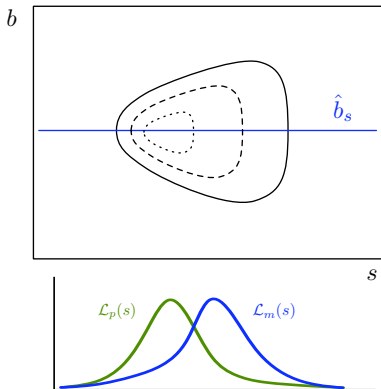
$$\begin{aligned}\mathcal{L}_m(s) &\equiv \int db \, p(b|s) \mathcal{L}(s, b) \\ &\approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s\end{aligned}$$


Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background *subtraction* is a special case of background *marginalization*.

Flared/skewed/bannana-shaped: \mathcal{L}_m and \mathcal{L}_p differ



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$. Otherwise, they will likely *differ*.

In *measurement error problems* the difference can be dramatic

The on/off problem for Poisson counting data

Basic problem

- Look off-source; unknown background rate b
Count N_{off} photons in interval T_{off}
- Look on-source; rate is $r = s + b$ with unknown signal s
Count N_{on} photons in interval T_{on}
- Infer s

Conventional solution

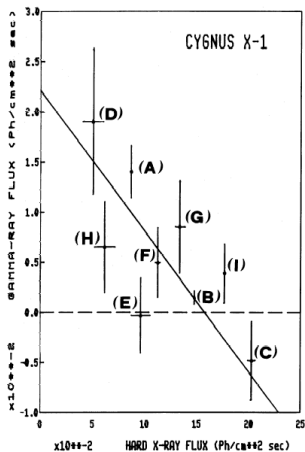
$$\begin{aligned}\hat{b} &= N_{\text{off}}/T_{\text{off}}; & \sigma_b &= \sqrt{N_{\text{off}}}/T_{\text{off}} \\ \hat{r} &= N_{\text{on}}/T_{\text{on}}; & \sigma_r &= \sqrt{N_{\text{on}}}/T_{\text{on}} \\ \hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2}\end{aligned}$$

But \hat{s} can be *negative*!

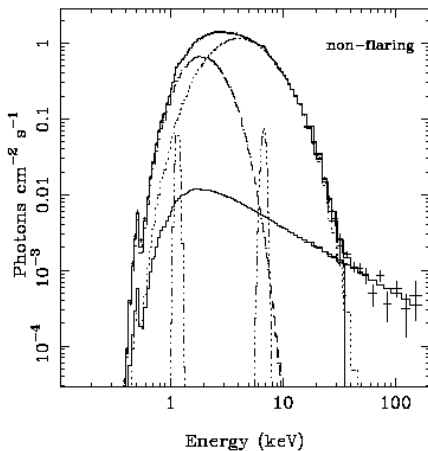
Examples

Spectra of X-ray sources

Bassani et al. 1989

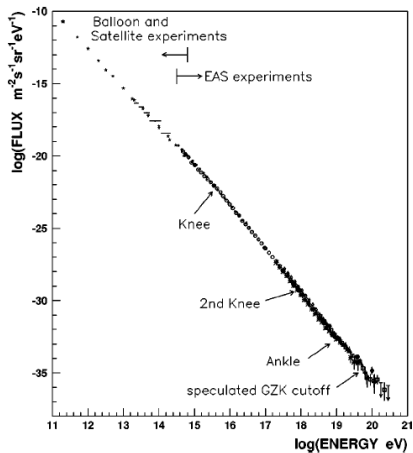


Di Salvo et al. 2001

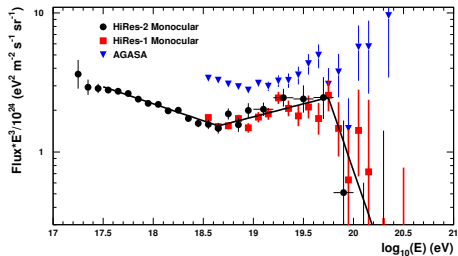


Spectrum of ultrahigh-energy cosmic rays

Nagano & Watson 2000



HiRes Team 2007



N is never large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)

N is never large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)

Bayesian solution to on/off problem

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

$$\mathcal{L}(s, b) = \frac{(bT_{\text{off}})^{N_{\text{off}}}}{N_{\text{off}}!} e^{-bT_{\text{off}}} \times \frac{[(s+b)T_{\text{on}}]^{N_{\text{on}}}}{N_{\text{on}}!} e^{-(s+b)T_{\text{on}}}$$

Adopting flat priors for (s, b) , the joint posterior is

$$p(s, b | N_{\text{on}}, N_{\text{off}}, \mathcal{C}) \propto (s+b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-sT_{\text{on}}} e^{-b(T_{\text{on}}+T_{\text{off}})}$$

Note if $b = 0$, the (normalized) posterior distribution is a gamma distribution,

$$p(s, b = 0 | N_{\text{on}}, N_{\text{off}}, \mathcal{C}) = \frac{T_{\text{on}}(sT_{\text{on}})^{N_{\text{on}}}}{N_{\text{on}}!} e^{-sT_{\text{on}}}$$

Now marginalize over b ;

$$\begin{aligned} p(s|N_{\text{on}}, N_{\text{off}}, \mathcal{C}) &= \int db \, p(s, b | N_{\text{on}}, \mathcal{C}) \\ &\propto \int db \, (s + b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-sT_{\text{on}}} e^{-b(T_{\text{on}} + T_{\text{off}})} \end{aligned}$$

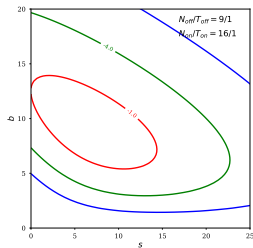
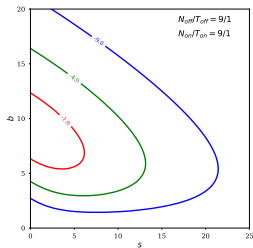
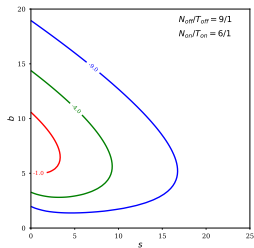
Expand $(s + b)^{N_{\text{on}}}$ and do the resulting Γ integrals:

$$\begin{aligned} p(s|N_{\text{on}}, N_{\text{off}}, \mathcal{C}) &= \sum_{i=0}^{N_{\text{on}}} C_i \frac{T_{\text{on}} (sT_{\text{on}})^i e^{-sT_{\text{on}}}}{i!} \\ C_i &\propto \left(1 + \frac{T_{\text{off}}}{T_{\text{on}}}\right)^i \frac{(N_{\text{on}} + N_{\text{off}} - i)!}{(N_{\text{on}} - i)!} \end{aligned}$$

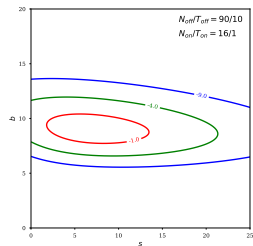
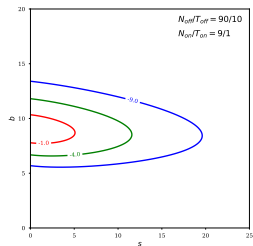
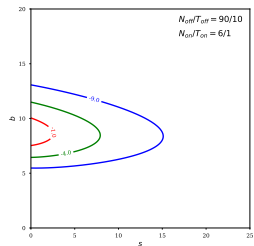
Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

Example on/off joint PDFs

$$T_{\text{on}} = T_{\text{off}} = 1$$

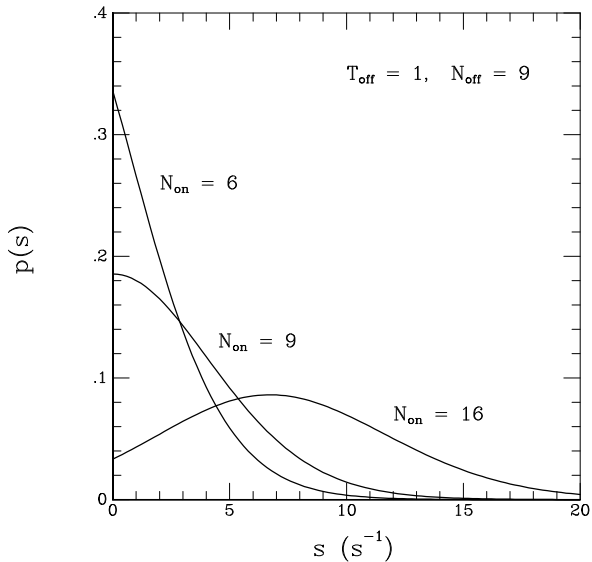


$$T_{\text{on}} = 1, T_{\text{off}} = 10$$



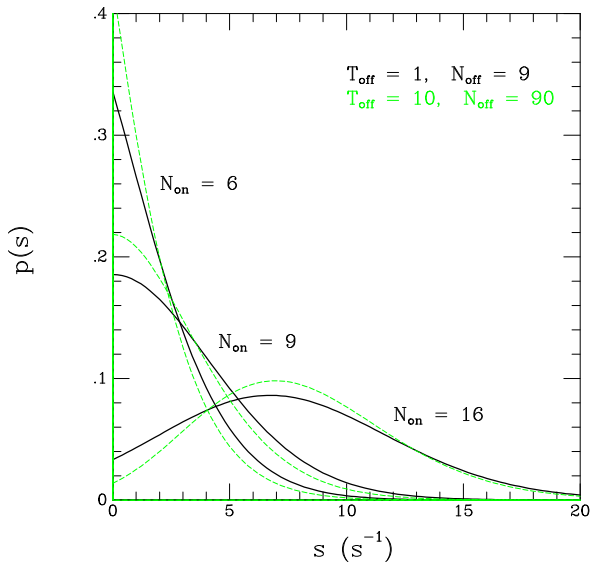
Example on/off marginal PDFs—Short integrations

$$T_{\text{on}} = T_{\text{off}} = 1$$



Example on/off marginal PDFs—Long background integrations

$$T_{\text{on}} = 1, T_{\text{off}} = 10$$



Supplement:

- Analytical details for Poisson dist'n inference
- Gamma-Poisson conjugate model
- Alternative (equivalent) solution to the on/off problem
- Multibin case

Many roles for marginalization

Eliminate nuisance parameters

$$p(\phi|D, M) = \int d\eta \, p(\phi, \eta|D, M)$$

Propagate uncertainty

Model has parameters θ ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F|D, M) &= \int d\theta \, p(F, \theta|D, M) = \int d\theta \, p(\theta|D, M) p(F|\theta, M) \\ &= \int d\theta \, p(\theta|D, M) \delta[F - f(\theta)] \quad [\textit{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D , predict future data D' (e.g., for *experimental design*):

$$p(D'|D, M) = \int d\theta \, p(D', \theta|D, M) = \int d\theta \, p(\theta|D, M) p(D'|\theta, M)$$

Model comparison. . .

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Supplement:

- Odds and Bayes factors: Compare models using *marginal* (average) likelihoods, not *maximum* likelihoods
- Bayesian Ockham's razor and Ockham factors
- Bayesian model averaging

Theme: Parameter space volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Model likelihoods have Ockham factors resulting from parameter space volume factors

Many virtues of Bayesian methods can be attributed to this accounting for the “size” of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added “by hand”—ignoring Fisher!).

On the key role of marginalization

Bayesian statistics uses all of probability theory, not just Bayes's theorem, and not even primarily Bayes's theorem. . . . Perhaps the most important theorem for doing Bayesian calculations is the *law of total probability* (LTP) that relates marginal probabilities to joint and conditional probabilities. . . . Arguably, if this approach to inference is to be named for a theorem, "total probability inference" would be a more appropriate appellation than "Bayesian statistics." It is probably too late to change the name. But it is not too late to change the emphasis.

— Loredó (2013)

The key distinguishing property of a Bayesian approach is marginalization instead of optimization, not the prior, or Bayes rule. . . . Broadly speaking, what makes Bayesian approaches distinctive is a posterior weighted marginalization over parameters. . . . Moreover, basic probability theory indicates that marginalization is desirable.

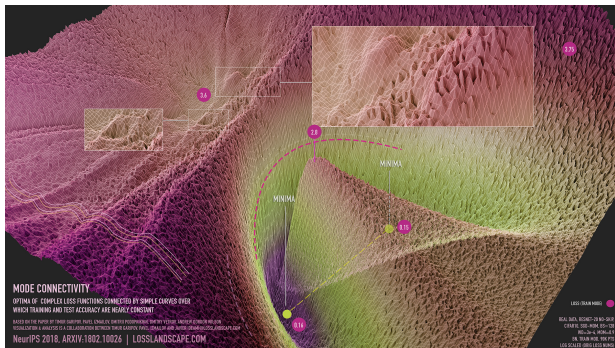
— Wilson (2020), Wilson & Izmailov (2020)

Marginalization can be hard: Bayesian neural nets

Neural nets are large, composite models with thousands to millions of weight parameters, w :

$$\begin{aligned}\log[p(w|D)] &= \log[\pi(w)] + \log[\mathcal{L}(w)] + C \\ &= \log[\pi(w)] - \text{Loss}(w) + C\end{aligned}$$

Deep neural net loss landscape



See: Loss surfaces... and What Are Bayesian Neural Network Posteriors Really Like?

Roles of the prior

Prior has two roles

- Modulate the likelihood to incorporate relevant prior information
- Convert likelihood from “intensity” to “measure”
→ enable accounting for *size of parameter space*

Physical analogy

$$\text{Heat } Q = \int d\vec{r} c_v(\vec{r}) T(\vec{r})$$

$$\text{Probability } P \propto \int d\theta p(\theta) \mathcal{L}(\theta)$$

Maximum likelihood focuses on the “hottest” parameters.

Bayes focuses on the parameters with the most “heat.”

A high- T region may contain little heat if its c_v is low or if its volume is small.

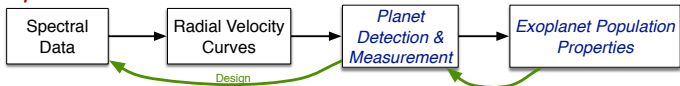
A high- \mathcal{L} region may contain little probability if its prior is low or if its volume is small.

Priors are like initial conditions/boundary conditions in physics: sometimes a nuisance, sometimes crucially important, always required by the theory

By converting likelihood to probability, priors provide two crucial capabilities:

- Accumulation of evidence (learning) → *discovery chains*

Exoplanets



- Automatic accounting for the sizes of parameter/hypothesis spaces: nuisance parameters, uncertainty propagation, prediction, model comparison. . .

If your problem needs particularly careful and thorough implementation of these capabilities, you should consider Bayesian methods

Supplement:

- Assigning priors
- Rule-based “objective” priors: Jeffreys, reference

Also see Stan’s “Prior Choice Recommendations” Wiki

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Bayesian curve fitting & least squares

Setup

Data $D = \{d_i\}$ are measurements of an underlying function $f(x; \theta)$ at N sample points $\{x_i\}$. Let $f_i(\theta) \equiv f(x_i; \theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2)$$

We seek to learn θ , or to compare different functional forms (model choice, M)

Likelihood

$$\begin{aligned} p(D|\theta, M) &= \prod_{i=1}^N \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right] \\ &\propto \exp \left[-\frac{1}{2} \sum_i \left(\frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right] \\ &= \exp \left[-\frac{\chi^2(\theta)}{2} \right] \end{aligned}$$

Posterior

For prior density $\pi(\theta)$,

$$p(\theta|D, M) \propto \pi(\theta) \exp \left[-\frac{\chi^2(\theta)}{2} \right]$$

If you have a least-squares or χ^2 code:

- Treat $\chi^2(\theta)$ as $-2 \log \mathcal{L}(\theta)$
- Bayesian inference amounts to exploration and *numerical integration* (by quadrature or Monte Carlo) of $\pi(\theta)e^{-\chi^2(\theta)/2}$

Forthcoming Python *Parametric Inference Engine* (PIE):

```
class MyData(PredictorSet):
    d1 = SampledGaussianPred(data1, doc="Sampled")
    d2 = BinnedGaussianPred(data2, doc="Binned")

class PowerLaw(SignalModel):
    A = PosParam(1., 'Amplitude')
    alpha = RealParam(range=(-5,-1), 'Index')

    def signal(self, E):
        return self.A * E**self.alpha
```

```
class PowerLawInference(BayesianInference,
                        PowerLaw, MyData):
    def log_prior(self):
        return 0. # const. prior

inf = PowerLawInference()

inf.A.vary()
inf.alpha.step(0., 5., 50)

grid1 = laplace() # Laplace approx.
grid2 = marg()    # Marg. via cubature
```

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

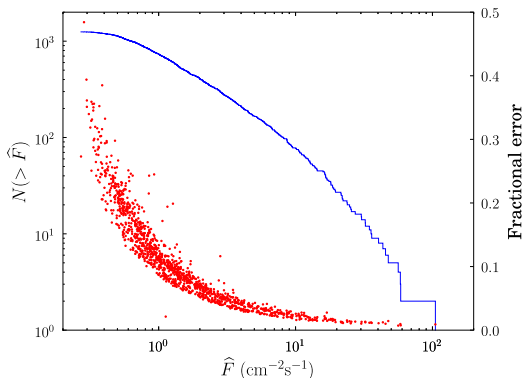
⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

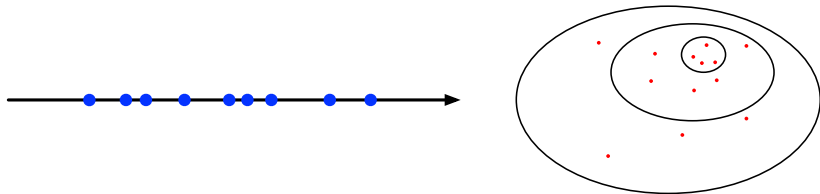
Motivation: Complications with survey data



- *Selection effects* (truncation, censoring) — *obvious* (usually)
Typically treated by “correcting” data
Most sophisticated: product-limit estimators
- *“Scatter” effects* (measurement error, etc.) — *insidious*
Typically ignored (average out? *No*—Eddington bias!)

Accounting for measurement error

Suppose $f(x|\theta)$ is a distribution for an observable, x (scalar or vector, $\vec{x} = (x, y, \dots)$); and θ is unknown



From N precisely measured samples, $\{x_i\}$, we can infer θ from

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

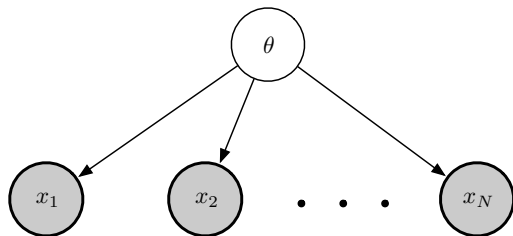
(A *binomial point process*)

$$p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\})$$

Posterior \propto joint for params & data

Graphical representation

- Nodes/vertices = uncertain quantities (gray \rightarrow known)
- Edges specify conditional dependence
- Absence of an edge denotes *conditional independence*

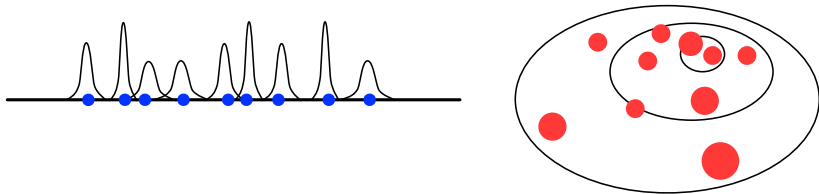


Graph specifies the form of the *joint distribution*:

$$p(\theta, \{x_i\}) = p(\theta) p(\{x_i\}|\theta) = p(\theta) \prod_i f(x_i|\theta)$$

Posterior from BT: $p(\theta|\{x_i\}) = p(\theta, \{x_i\})/p(\{x_i\})$

But what if the x data are *noisy*, $D_i = \{x_i + \epsilon_i\}$?



$\{x_i\}$ are now *uncertain (latent/hidden/incidental) parameters*

Member/item likelihoods quantify uncertainties: $\ell_i(x_i) = p(D_i|x_i)$

The joint PDF for *everything* is

$$\begin{aligned} p(\theta, \{x_i\}, \{D_i\}) &= p(\theta) p(\{x_i\}|\theta) p(\{D_i\}|\{x_i\}) \\ &= p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i) \end{aligned}$$

The conditional (posterior) PDF for the unknowns is

$$p(\theta, \{x_i\}|\{D_i\}) = \frac{p(\theta, \{x_i\}, \{D_i\})}{p(\{D_i\})} \propto p(\theta, \{x_i\}, \{D_i\})$$

$$\begin{aligned}
 p(\theta, \{x_i\} | \{D_i\}) &\propto p(\theta, \{x_i\}, \{D_i\}) \\
 &= p(\theta) \prod_i f(x_i | \theta) \ell_i(x_i)
 \end{aligned}$$

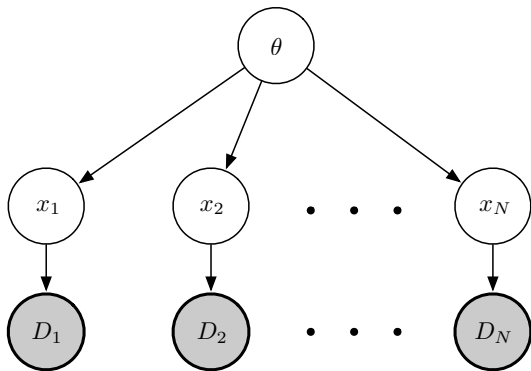
Marginalize over $\{x_i\}$ to summarize inferences for θ

Marginalize over θ to summarize inferences for $\{x_i\}$

Key point: *Maximizing over x_i (i.e., just using best-fit/MLE \hat{x}_i) and integrating over x_i can give very different results!*

(See Loredó (2004) for tutorial examples)

Graphical representation



$$\begin{aligned} p(\theta, \{x_i\}, \{D_i\}) &= p(\theta) p(\{x_i\}|\theta) p(\{D_i\}|\{x_i\}) \\ &= p(\theta) \prod_i f(x_i|\theta) p(D_i|x_i) = p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i) \end{aligned}$$

A two-level *hierarchical Bayes model*, *multi-level model* (MLM), or *probabilistic graphical model* (PGM)

Hierarchical Bayes/MLMs/PGMs in astronomy

A few entry points (see the *Resources* document for books with HB content):

- Joel Leja's tutorial at the 2022 Astrominformatics Summer School (next Wednesday!)
- Chapter by Loredó & Hendry: Multilevel and hierarchical Bayesian modeling of cosmic populations
- Survey of MLMs in astronomy: Bayesian astrostatistics: A backward look to the future (TJL 2013)
- CAST 2014 Supplement Session — Includes discussion of selection effects
- TL's AAS workshop at AAS 231 (2018): Hierarchical modeling of cosmic populations
- CUDAHM C++ GPU software (Szalai-Gindl, Budavari, Kelly, TL); see Astron. & Comp. paper (longer: arXiv:2105.08026)

Agenda

① Quantifying uncertainty with probability

② Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$ via Monte Carlo

Confidence intervals vs. credible intervals

③ Probability theory for data analysis: Three theorems

④ Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

⑤ Quick-looks

Curve fitting & least squares

Measurement error & hierarchical/graphical models

Bayesian computation menu (1 slide!)

Bayesian computation menu

Large sample size, N : Laplace approximation

- Approximate posterior as multivariate normal $\rightarrow \det(\text{covar})$ factors
- Uses ingredients available in χ^2 /ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

Modest-dimensional models ($m \lesssim 10$ to 20)

- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo

- Posterior sampling — create RNG that samples posterior
 - ▶ Markov Chain Monte Carlo (MCMC) is the most general framework
- Nested sampling
- Sequential Monte Carlo (SMC)
- Approximate(ly) Bayesian computation (ABC)/Likelihood-free inference (LFI)
- ...

Recap of key ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis \propto ability of hypothesis to *predict* the observed data

Law of total probability

$$p(\text{Hypotheseses} \mid \text{Data}) = \sum p(\text{Hypothesisis} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true